A Mood Analysis on Youtube Comments and a Method for Improved Social Spam Detection

Enaitz Ezpeleta, Mikel Iturbe, Iñaki Garitano, Iñaki Velez de Mendizabal, and Urko Zurutuza

Electronics and Computing Department, Mondragon University
Goiru Kalea, 2, 20500 Arrasate-Mondragón, Spain
{eezpeleta,miturbe,igaritano,ivelez,uzurutuza}@mondragon.edu

Abstract. In the same manner that Online Social Networks (OSN) usage increases, non-legitimate campaigns over these types of web services are growing. This is the reason why significant number of users are affected by social spam every day and therefore, their privacy is threatened. To deal with this issue in this study we focus on mood analysis, among all content-based analysis techniques. We demonstrate that using this technique social spam filtering results are improved. First, the best spam filtering classifiers are identified using a labeled dataset consisting of Youtube comments, including spam. Then, a new dataset is created adding the mood feature to each comment, and the best classifiers are applied to it. A comparison between obtained results with and without mood information shows that this feature can help to improve social spam filtering results: the best accuracy is improved in two different datasets, and the number of false positives is reduced 13.76% and 11.41% on average. Moreover, the results are validated carrying out the same experiment but using a different dataset.

Keywords: spam, social spam, mood analysis, online social networks, Youtube

1 Introduction

In recent years, Online Social Networks (OSNs) have extensively expanded around the world. The amount of users per each OSN platform shows the importance of these communication channels in our society: Facebook reached 1.4 billion daily active users on average as of December 2017 [1]; Youtube has counted over a billion users in 2017 [2]; and Twitter has 330 million monthly active users as of June 30, 2017[3].

The sudden increase in users gives malicious organizations the possibility to reach a vast amount of people easily. Authors in [11] demonstrate that these

[1] http://newsroom.fb.com/company-info/
[2] https://www.youtube.com/yt/about/press/
[3] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

sites are now a major delivery platform targeted for spam. They analyze spam in several OSNs, and they quantify spam campaigns delivered from accounts in OSNs.

To deal with this problem, authors in [8] prove that content-based techniques can help to improve spam filtering results. They perform sentiment analysis on email messages in order to enrich the original dataset, and they obtain improved accuracy. Following a similar procedure, in this study we use another content-based technique, the mood analysis of the messages, to improve social media spam filtering results.

First, several spam filtering classifiers and different settings are applied to a Youtube comment dataset in order to identify the best ten filtering methods. After that, a mood analyzer is applied to each comment with the purpose of creating a new dataset adding this feature to the original dataset. Once the enhanced dataset is created, the previously selected ten classifiers are applied to the dataset with the mood feature. Finally, a comparison and an analysis of the results is carried out.

The remainder of this paper is organized as follows. Section 2 describes the previous work conducted in the area of social media spam filtering techniques. Section 3 describes the process of the aforementioned experiments, regarding Bayesian spam filtering and spam filtering using the mood of the texts. In Section 4, the obtained results are described, and finally, we summarize our findings and give conclusions in Section 5.

2 Related Work

OSN-related spam is an active research field [2] that has received wide attention from the scientific community. Stringhini et al. [22] demonstrated that it is possible to identify spammer accounts in large OSNs in an automatic manner and later block them, applying their approach in Facebook, Twitter and MySpace. Moreover, Wang et al. [25] proposed a spam detection system that is able to analyze OSNs in search of spam. Similarly, Egele et al. [7] presented COMPA, a tool to detect compromised OSN accounts based on anomalous user behavior. In a similar note, Gao et al. [10] employed classification and clustering to detect OSN spam campaigns in Twitter and Facebook. Ezpeleta et al. [9] showed that personalizing spam messages using publicly available OSN profile information lead to a significantly higher success rate than conventional, non-personalized spam.

In the field of OSN spam, the case of Twitter-based spam has been particularly well studied. Kwak et al. [13] identify this OSN as a useful tool for information diffusion. Therefore, it can be stated that Twitter is an attractive platform to perform spam campaigns. Yang et al. [26] described the dynamics of criminal accounts in Twitter and how they interact between them, and to other non-criminal accounts. Additionally, Song et al. [21] showed how to detect spammers based on the measurement of relation features, such as the distance and connectivity between receiver and recipient, instead of focusing on Twitter

account features, as these account features are more prone to spammer manipulation.

Even if studied not as much as email, Twitter or Facebook, Youtube spam has also been an object of study. Chaudhary and Sureka [6] mined video descriptions, along with temporal and popularity based features, to detect spam videos on Youtube. O'Callaghan et al. [15] use network motif profiling to identify recurring Youtube spam campaigns, by characterizing Youtube users as motifs. By identifying users with distinctive motifs, they where able to label users in spamming campaigns.

However, even if numerous novel spam detection techniques have been published [24, 27, 20], OSN spam messages remain an open problem yet to be solved [11].

In this direction, content-based analysis, where text is analyzed to infer its meaning or purpose using different techniques such as Sentiment Analysis (SA), stands as a promising procedure for improving spam detection in OSNs [19, 8]. The main objective of SA resides in the identification of the positive or negative nature of a document [17]. In order to reach this objective, it is possible to use a supervised learning approach, with three previously defined classes (positive, negative and neutral) [18] or a unsupervised one, where opinion words or phrases are the dominating indicators for sentiment classification [23].

It has been already demonstrated that SA helps in spam detection in different cases, such as social spammer detection [12], short informal messages [3], Twitter spam [19], email spam [8] and fraud detection [14].

As authors presented in [5], sentiment extraction from text can be used to predict mood, which can be used to prevent or mitigate security threats. Defined as "a temporary state of mind or feeling"[4], mood was used by Bollen et al [4] in their analysis of Twitter feeds.

In this paper, we go beyond the State of the Art by using mood analysis for improved spam filtering, focusing on a popular OSN, the Youtube video service. To the best of our knowledge, no previous research has focused on using this approach for improved spam detection, let alone in the case of analyzing Youtube comments.

3 Design and Implementation

Having an original dataset, the process followed in this study is divided in two main parts, depicted on Figure 1.

1. First, several classifiers are applied to a dataset consisting of social media messages (spam and ham) in order to identify and select the best ten social spam filtering classifiers. In the same step, the best 10 results are also extracted.
2. Second, the mood of each message is added to the original dataset to create a new dataset. During the mood analysis, a descriptive experiment is carried

[4] https://en.oxforddictionaries.com/definition/mood

out. In the next phase, the best ten classifiers selected in the previous step are applied to the created dataset in order to compare the results.

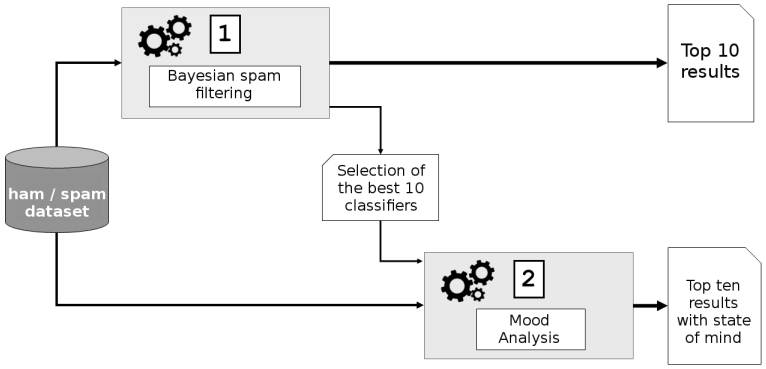

Fig. 1. Improving social spam detection using the mood of the comments.

To validate the algorithms and the obtained results the 10-fold cross-validation technique is used, and the results are analyzed in terms of the number of false positives and the accuracy. Accuracy is the percentage of testing set examples correctly classified by the classifier. Legitimate messages classified as spam are considered false positives. In order to validate these results, the same process is followed using another dataset.

3.1 Datasets

During this work two publicly available datasets are used:

- *Youtube Comments Dataset* [5]: Presented in [16]. This dataset contains multilingual 6,431,471 comments from a popular social media website, Youtube[6]. Among all the comments, 481,334 are marked as spam.
  In order to use similar number of texts messages to the experiments presented in [8] we created a new subset consisting of 1,000 spam and 3,000 ham, i.e. legitimate, comments. Those texts have been selected randomly and only taking into account comments written in English.
- *YouTube Spam Collection Dataset* [7]: Published by Alberto et al. [1]. Composed by 1,956 real messages divided in five subsets. The comments were extracted from five out of the ten most popular videos on the collection period. It consists of 1,005 spam and 951 ham texts. During this study, we use this dataset to validate the results of the previous dataset, repeating the experimental workflow.

[5] http://mlg.ucd.ie/yt/
[6] www.youtube.com
[7] https://archive.ics.uci.edu/ml/datasets/YouTube+Spam+Collection

3.2 Identifying the best social spam classifiers

With the objective of identifying the best spam detectors, several spam classifiers using different settings are applied to the Youtube Comments dataset.

Following the strategy presented in [8], 7 different classifiers and 56 settings combinations per each classifier are applied (we apply 392 combinations in total) The best ten results are presented in Table 1. During this experiment seven different classifiers have been used: (1) Large-scale Bayesian logistic regression for text categorization, (2) discriminative parameter learning for Bayesian networks, (3) Naive Bayes classifier, (4) complement class Naive Bayes classifier, (5) multi-nominal Naive Bayes classifier, (6) updateable Naive Bayes classifier, and (7) updateable multi-nominal Naive Bayes classifier.

Table 1. Results of the best ten classifiers

#	Spam classifier	TP	TN	FP	FN	Acc
1	NBM.c.stwv.go.ngtok	389	2911	89	611	82.50
2	NBMU.c.stwv.go.ngtok	389	2911	89	611	82.50
3	NBM.stwv.go.ngtok	370	2929	71	630	82.48
4	NBMU.stwv.go.ngtok	370	2929	71	630	82.48
5	NBM.c.stwv.go.ngtok.stemmer	379	2919	81	621	82.45
6	NBMU.c.stwv.go.ngtok.stemmer	379	2919	81	621	82.45
7	NBM.stwv.go.ngtok.stemmer	358	2936	64	642	82.35
8	NBMU.stwv.go.ngtok.stemmer	358	2936	64	642	82.35
9	CNB.stwv.go.ngtok	417	2875	125	583	82.30
10	CNB.stwv.go.ngtok.stemmer	400	2891	109	600	82.28

Nomenclatures and acronyms used in Table 1 and also throughout the paper are explained in Table 2.

Table 2. Nomenclatures

	Meaning
CNB	Complement Naive Bayes
NBM	Naive Bayes Multinomial
NBMU	Naive Bayes Multinomial Updatable
.c	idft F, tft F, outwc T [8]
.i.c	idft T, tft F, outwc T [8]
.i.t.c	idft T, tft T, outwc T [8]

	Meaning
.stwv	String to Word Vector
.go	General options
.wtok	Word Tokenizer
.ngtok	NGram Tokenizer 1-3
.stemmer	Stemmer
.igain	Attribute selection using InfoGainAttributeEval

[8] idft means Inverse Document Frequency (IDF) Transformation; tft means Term Frequency score (TF) Transformation; outwc counts the words occurrences.

Once the best classifiers and the best results are identified using the Youtube Comments dataset, a mood analysis of each message is carried out.

3.3 Mood analysis

In order to analyze the mood of the youtubers' comments, each text is analyzed and a new feature (mood) is added to the original dataset. In this way a new dataset is created, and the best ten classifiers identified in the first phase are applied to it. This process is presented in Figure 2.

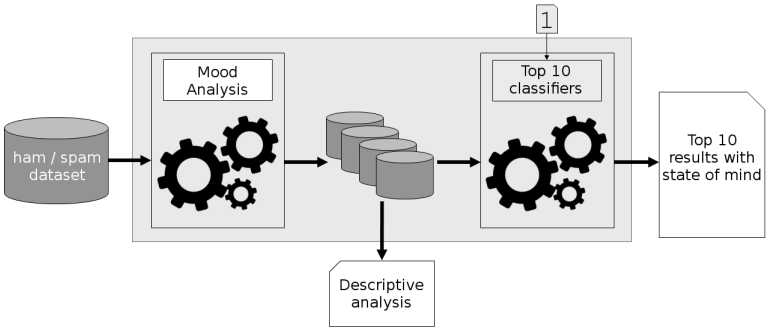

Fig. 2. Mood analysis.

To extract the mood of the writer, we use a publicly available machine learning Software as a Service (SaaS). This tool is hosted in *uClassify*[9], and taking into account the reached accuracy (96%) among all the possibilities we selected Mood classifier developed by Mattias Östmar.

As the author explains, this function determines the state of mind of the writer (upset or happy). On the extreme side there are angry, hateful writers while on the other extreme there are joyful and loving writers. The accuracy of the presented analyzer was measured using 10-fold cross validation by the author, and a 96% was reached.

The web service returns a float within the range [0.0, 1.0] specifying the happiness of each text. Using this value, it is possible to calculate also the upset level. Consequently, only one feature (mood) per each comment is added to the original dataset, and a new dataset is created.

4 Experimental Results

In order to achieve the objective of this study, several experiments are carried out using the previously mentioned Youtube comments dataset. The results of these tests are presented in this section.

[9] https://uclassify.com

4.1 Descriptive analysis

First, we perform a descriptive experiment of the two publicly available datasets in terms of the mood of the comments. We start by applying the mood analyzer to the dataset, then we continue by extracting statistics about the distribution and finally, we add a new feature, mood, to the original dataset.

As a result of this analysis, we find out that the state of mind of the spam texts in Youtube differs depending on the data collection strategy. On the one hand, Youtube Comments Dataset was created by crawling the comments from 6,407 different videos. On the other hand, researchers used only 5 out of the 10 most popular videos on the collection period to create the YouTube Spam Collection Dataset. The difference between datasets and also between ham and spam message is shown in the box plot presented in Figure 3.

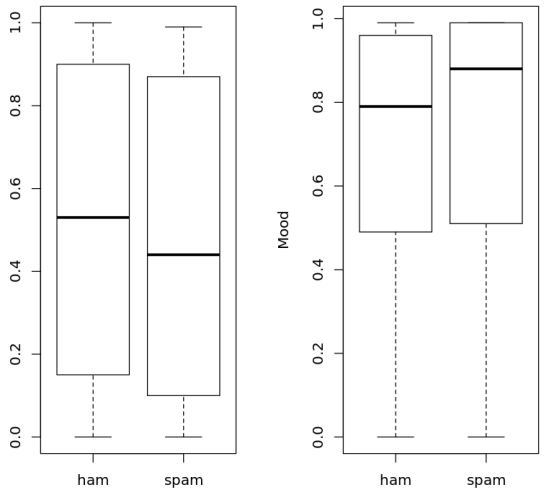

Fig. 3. Mood analysis of the dataset. (left-hand side) Youtube Comments Dataset and (right-hand side) YouTube Spam Collection Dataset.

This means that there is a difference between spam and ham social comments in terms of mood, so so the discriminative nature of this feature can aid in improving social spam filtering.

4.2 Predictive experiment

To analyze the influence of the mood analysis in social spam filtering a predictive experiment is carried out.

We apply the best ten classifiers identified in the Youtube Comments Dataset and we compare the results with and without mood feature. The comparison between different results is presented in Table 3.

Table 3. Comparison between the best ten classifiers with and without mood. Using the first dataset.

	Normal		*Mood*		
Name	**FP**	**Acc**	**FP**	**Acc**	**FP reduction (%)**
NBM.c.stwv.go.ngtok	89	82.50	**77**	**82.53**	13.48
NBMU.c.stwv.go.ngtok	89	82.50	**70**	**82.58**	21.35
NBM.stwv.go.ngtok	71	**82.48**	**63**	82.43	11.27
NBMU.stwv.go.ngtok	71	**82.48**	**59**	82.43	16.90
NBM.c.stwv.go.ngtok.stemmer	81	82.45	**73**	**82.45**	9.88
NBMU.c.stwv.go.ngtok.stemmer	81	82.45	**68**	**82.48**	16.05
NBM.stwv.go.ngtok.stemmer	64	82.35	**58**	**82.38**	9.38
NBMU.stwv.go.ngtok.stemmer	64	82.35	**53**	**82.35**	17.19
CNB.stwv.go.ngtok	125	82.30	**110**	**82.43**	12.00
CNB.stwv.go.ngtok.stemmer	109	**82.28**	**98**	82.20	10.09
				avg:	*13.76*

As it is possible to see in Table 3, the comparison has been done taking into account the accuracy and the number of false positives. Results show that in almost all the top classifiers the accuracy is improved with the added mood analysis, and the best accuracy is obtained reaching an 82.58%. Moreover, the number of false positives is reduced in every cases between 9.38% and 21.35%.

In order to validate the results obtained with the Youtube Comments Dataset, the same experiments are carried out using the YouTube Spam Collection Dataset. The obtained results are shown in Table 4.

Table 4. Comparison between the best ten classifiers with and without mood. Using the validation dataset.

	Normal		*Mood*		
Name	**FP**	**Acc**	**FP**	**Acc**	**FP reduction (%)**
CNB.stwv.go.ngtok	85	93.97	**76**	**94.38**	10.59
CNB.stwv.go.ngtok.stemmer	89	93.87	**85**	**94.02**	4.49
NBM.stwv.go.ngtok	113	92.69	**80**	**94.17**	29.20
NBMU.stwv.go.ngtok	**113**	**92.69**	116	92.54	-2.65
NBM.stwv.go.ngtok.stemmer	119	92.38	**86**	**93.97**	27.73
NBMU.stwv.go.ngtok.stemmer	**119**	**92.38**	123	92.23	-3.36
NBM.c.stwv.go.ngtok	127	92.13	**96**	**93.66**	24.41
NBMU.c.stwv.go.ngtok	127	92.13	**127**	**92.13**	0.00
NBM.c.stwv.go.ngtok.stemmer	135	91.72	**101**	**93.35**	25.19
NBMU.c.stwv.go.ngtok.stemmer	**135**	**91.72**	137	91.62	-1.48
				avg:	*11.41*

Using the validation dataset results are also improved adding mood feature. In this case the best accuracy is improved from 93.97% to 94.38%, and the number of false positives is reduced 11.41% on average.

Furthermore, to compare statistical behavior of the best classifier in both datasets, the area under the curve is analyzed. To create this ROC curve the specificity and sensitivity of the classifiers are taken into account. Figure 4 shows that the ROC area using mood analysis (0.756 and 0.943) is larger than without using it (0.753 and 0.939) in both datasets.

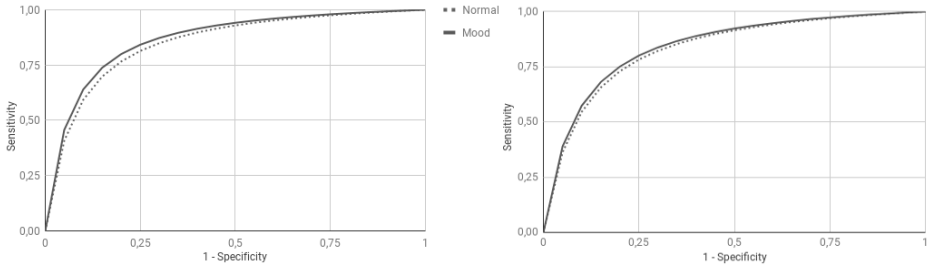

Fig. 4. ROC Curve of the best classifier with and without mood. (left-hand side) Youtube Comments Dataset and (right-hand side) YouTube Spam Collection Dataset.

5 Conclusions

This paper presents a new social spam filtering method. We provide means to validate our hypothesis that it is possible to improve current social spam filtering results extracting the mood of the texts.

First, using a social spam dataset, different experiments are carried out with and without mood feature. Next, we compare the obtained results, and we demonstrate that mood analysis can help to improve social spam filtering results.

Results show that the best accuracy obtained with the original dataset is improved from 82.50% to 82.58% using the Youtube Comments Dataset, and from 93.97% to 94.38% using the validation dataset. Despite the difference in the percentage does not seem to be relevant, if we take into account the amount of Youtube comments and the daily active users in this website, the improvement in absolute comment number is significant. Additionally, the number of false positives is reduced, on average 13.76% and 11.47%. This means that mood analysis is capable to highlight differences between spam and legitimate social comments. As descriptive analysis shows, the mood feature adds a distinctive feature for comments in each type of video (more positive or more upset). This variation helps classifiers to filter spam comments, and to improve the results.

Acknowledgments. This work has been developed by the intelligent systems for industrial systems group supported by the Department of Education, Language policy and Culture of the Basque Government. This work was partially supported by the project Semantic Knowledge Integration for Content-Based Spam Filtering (TIN2017-84658-C2-2-R) from the Spanish Ministry of Economy, Industry and Competitiveness (SMEIC), State Research Agency (SRA) and the European Regional Development Fund (ERDF).

We thank Mattias Östmar for the valuable tools developed and published. And we thank Jon Kâgström (Founder of uClassify[10]) for the opportunity to use their API for research purposes.

References

1. Alberto, T.C., Lochter, J.V., Almeida, T.A.: Tubespam: Comment spam filtering on youtube. In: 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA). pp. 138–143 (Dec 2015)
2. Almaatouq, A., Shmueli, E., Nouh, M., Alabdulkareem, A., Singh, V.K., Alsaleh, M., Alarifi, A., Alfaris, A., Pentland, A.S.: If it looks like a spammer and behaves like a spammer, it must be a spammer: analysis and detection of microblogging spam accounts. International Journal of Information Security 15(5), 475–491 (2016), http://dx.doi.org/10.1007/s10207-016-0321-5
3. Arif, M.H., Li, J., Iqbal, M., Liu, K.: Sentiment analysis and spam detection in short informal text using learning classifier systems. Soft Computing (Jul 2017), https://doi.org/10.1007/s00500-017-2729-x
4. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of computational science 2(1), 1–8 (2011)
5. Chandramouli, R.: Emerging social media threats: Technology and policy perspectives. In: 2011 Second Worldwide Cybersecurity Summit (WCS). pp. 1–4 (June 2011)
6. Chaudhary, V., Sureka, A.: Contextual feature based one-class classifier approach for detecting video response spam on youtube. In: 2013 Eleventh Annual Conference on Privacy, Security and Trust. pp. 195–204 (July 2013)
7. Egele, M., Stringhini, G., Kruegel, C., Vigna, G.: Compa: Detecting compromised accounts on social networks. In: NDSS. The Internet Society (2013), http://dblp.uni-trier.de/db/conf/ndss/ndss2013.html#EgeleSKV13
8. Ezpeleta, E., Zurutuza, U., Gómez Hidalgo, J.M.: Does sentiment analysis help in bayesian spam filtering? In: Hybrid Artificial Intelligent Systems: 11th International Conference, HAIS 2016, Sevilla, Spain, April 18-20, 2016. Springer (2016)
9. Ezpeleta, E., Zurutuza, U., Hidalgo, J.M.G.: A study of the personalization of spam content using facebook public information. Logic Journal of the IGPL 25(1), 30–41 (2017), http://dx.doi.org/10.1093/jigpal/jzw040
10. Gao, H., Chen, Y., Lee, K., Palsetia, D., Choudhary, A.N.: Towards online spam filtering in social networks. In: NDSS. The Internet Society (2012), http://dblp.uni-trier.de/db/conf/ndss/ndss2012.html#GaoCLPC12
11. Gao, H., Hu, J., Wilson, C., Li, Z., Chen, Y., Zhao, B.Y.: Detecting and characterizing social spam campaigns. In: Proceedings of the 17th ACM conference on

[10] https://www.uclassify.com

Computer and communications security. pp. 681–683. CCS '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1866307.1866396
12. Hu, X., Tang, J., Gao, H., Liu, H.: Social spammer detection with sentiment information. In: Proceedings of the 2014 IEEE International Conference on Data Mining. pp. 180–189. ICDM '14, IEEE Computer Society, Washington, DC, USA (2014), http://dx.doi.org/10.1109/ICDM.2014.141
13. Kwak, H., Lee, C., Park, H., Moon, S.: What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on World wide web. pp. 591–600. ACM (2010)
14. Mahajan, S., Rana, V.: Spam detection on social network through sentiment analysis. Advances in Computational Sciences and Technology 10(8), 2225–2231 (2017)
15. O'Callaghan, D., Harrigan, M., Carthy, J., Cunningham, P.: Network analysis of recurring youtube spam campaigns. arXiv preprint arXiv:1201.3783 (2012)
16. O'Callaghan, D., Harrigan, M., Carthy, J., Cunningham, P.: Network analysis of recurring youtube spam campaigns. CoRR abs/1201.3783 (2012), http://arxiv.org/abs/1201.3783
17. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval 2(1-2), 1–135 (2008)
18. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: Sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10. pp. 79–86. EMNLP '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), http://dx.doi.org/10.3115/1118693.1118704
19. Perveen, N., Missen, M.M.S., Rasool, Q., Akhtar, N.: Sentiment based twitter spam detection. International Journal of Advanced Computer Science and Applications(IJACSA) 7(7), 568–573 (2016)
20. Shehnepoor, S., Salehi, M., Farahbakhsh, R., Crespi, N.: Netspam: a network-based spam detection framework for reviews in online social media. IEEE Transactions on Information Forensics and Security PP(99), 1–1 (2017)
21. Song, J., Lee, S., Kim, J.: Spam filtering in twitter using sender-receiver relationship. In: Recent Advances in Intrusion Detection. pp. 301–317. Springer (2011)
22. Stringhini, G., Kruegel, C., Vigna, G.: Detecting spammers on social networks. In: Proceedings of the 26th Annual Computer Security Applications Conference. pp. 1–9. ACSAC '10, ACM, New York, NY, USA (2010), http://doi.acm.org/10.1145/1920261.1920263
23. Turney, P.D.: Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 417–424. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), http://dx.doi.org/10.3115/1073083.1073153
24. Wang, A.H.: Don't follow me: Spam detection in twitter. In: Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on. pp. 1–10. IEEE (2010)
25. Wang, D., Irani, D., Pu, C.: A social-spam detection framework. In: Proceedings of the 8th Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference. pp. 46–54. ACM (2011)
26. Yang, C., Harkreader, R., Zhang, J., Shin, S., Gu, G.: Analyzing spammers' social networks for fun and profit: a case study of cyber criminal ecosystem on twitter. In: Proceedings of the 21st international conference on World Wide Web. pp. 71–80. ACM (2012)

27. Zheng, X., Zeng, Z., Chen, Z., Yu, Y., Rong, C.: Detecting spammers on social networks. Neurocomputing 159, 27 – 34 (2015), http://www.sciencedirect.com/science/article/pii/S0925231215002106